

Overview of Student Loan Repayment Rate: A Comparative Study Implementing ANOVA, Machine Learning, and Ensemble Methods

Prabha Shrestha ^{a*}

Sarbagya Ratna Shakya ^a

Paola Selene Pereda ^a

^{a)} Department of Mathematical Sciences, Eastern New Mexico University, Portales, NM, USA

ABSTRACT

With rising student loan debt in the United States and a challenging job market, the risk of loan default poses a significant threat to graduates' financial futures. Currently, one in four U.S. adults under 40 carries student loan debt, totaling approximately \$1.73 trillion as of mid-2024. This study aims to predict trends in federal student loan repayment statuses and trends among students who completed their degrees within six years, assessing whether they remain in repayment seven years post-graduation, using machine learning techniques. Utilizing institution-level data from the U.S. Department of Education's College Scorecard (2013–2017), we analyzed fourteen key predictor variables, including six-year completion rates, student income, gender, and Pell Grant status. Seven machine learning models were implemented—Random Forest (RF), Gradient Boosting, Decision Tree, Bagging Regressor, K-Nearest Neighbors, eXtreme Gradient Booster (XGB) Regressor, and Nu Support Vector Regressor (NuSVR). The top-performing models (RF, XGB, and NuSVR) were integrated through an ensemble approach to enhance prediction accuracy. Results indicate that machine learning models can effectively predict repayment rates, with ensemble approaches enhancing performance as measured by Root Mean Squared Error (RMSE) and R-squared (R^2) metrics. Additionally, Analysis of Variance (ANOVA) analysis revealed statistically significant differences in repayment rates across the 2013–2017 cohorts ($p = 1.87E-62$). These findings demonstrate the potential of data-driven approaches to provide valuable insights into student loan repayment behaviors, informing policies that aim to improve financial security for graduates.

KEYWORDS: Federal loans, Machine learning methods, Financial futures, Data science, ANOVA, Tukey-Kramer, Students' repayment rate, Ensemble methods

1 INTRODUCTION

Student loan debt refers to the money borrowed through federal or private financial lenders to cover educational expenses, including tuition fees, living costs, books, and other educational necessities that help fill financial gaps (Diaz, 2025). The rising cost of tuition and the pursuit of stable, well-paying employment post-graduation have compelled many students in the United States to rely heavily on federal and private student loans (Melton, 2023). These loans provide critical financial support, especially for students who lack familial financial resources, enabling them to pursue and continue their higher education. However, this access comes with long-term risks

when borrowers cannot meet academic and financial expectations (Jabbari et al., 2023). Consequences include difficulties repaying loans after graduation and the admission of academically weaker students due to easier college access via loans. Consequently, student loan debt has become a major national issue that has drawn increasing attention in recent years (Heller, 2008).

For decades, the U.S. government has provided financial assistance to students pursuing higher education. Over the past 20 years, student loan debt has more than doubled, coinciding with a sharp rise in college enrollment. Today, one in four U.S. adults under the age of 40 carries student loan debt. As of the second quarter of 2024, total student loan debt in the United States is approximately \$1.73 trillion, comprising 92.8% in federal loans (\$1.60 trillion) and 7.2% in private loans (\$125 billion). Notably, 2023 marked the first annual decrease in student loan debt, with a decline of 0.96%. Despite this slight drop, federal student loan debt has grown at an average annual rate of 13.2% since 2007. Currently, 43.2 million Americans hold federal student loan debt. As of January 25, 2025, the average federal student loan balance per borrower is \$38,375, while the combined average balance, including private loans, reaches as high as \$41,520. Many borrowers may take up 20 years to repay their student loans in full (Willeford, 2025).

In 2023, several policies were introduced by the Biden presidential administration aimed at reducing payments and expanding access to federal student loan forgiveness. The Saving on a Valuable Education (SAVE) relief plan was blocked by a federal appeals court on August 9, 2024 (Hegji and Stiff, 2024). Despite legal challenges, these programs successfully canceled nearly \$190 billion in federal student loan debt for approximately 5.3 million borrowers. However, many federal student loan programs have encountered legal challenges, resulting in ongoing uncertainty regarding repayment plans. While millions have benefited from loan forgiveness, the policy has not impacted all federal loan borrowers. Critics argue that loan forgiveness policies may encourage over-borrowing and contribute to rising college costs (Edwards, 2016). Given the significant negative impacts of student debt on many borrowers, including its potential effects on the U.S. economy, student loan debt remains a vital and emerging issue that warrants focused study.

The literature suggests student loan debt and repayment outcomes are shaped by multiple factors in addition to tuition, including living expenses, institution type, enrollment rates, gender, race, and post-graduation financial management. For example, borrowers who attended private, non-profit colleges generally owe more than those who attended public non-profit institutions (10 Key Facts about Student Debt in the United States, n.d.). These factors cumulatively impact both individual financial health and the broader economy over the long term (10 Key Facts about Student Debt in the United States, n.d.). Strategies to minimize student debt include attending public in-state universities or community colleges with transfer pathways, taking on-campus part-time jobs such as a resident assistant position to offset housing and food costs, actively seeking scholarships, maintaining open communication with advisors, budgeting carefully, and exploring summer programs (Edwards, 2016). Thoughtful career planning post-graduation can further improve a student's ability to repay loans in a timely manner (Kuan et al., 2025).

The objective of this paper is to enhance understanding of the financial realities students face after graduation, with an emphasis on loan repayment status. Specifically, this study aims to analyze trends and significant differences in student loan repayment rates across multiple years, assess whether machine learning (ML) techniques can identify statistically significant relationships in repayment patterns among cohorts of federal student loan borrowers, and predict repayment rates for future student loans, focusing on federal loan borrowers who attended public institutions and completed their programs within six years.

This paper is organized as follows: Section 2 reviews relevant literature, covering prior research, proposed repayment solutions, and applications of ML and statistical techniques. Section 3 describes the dataset and data cleaning processes. Section 4 outlines the ML methods employed and presents the analyses. Section 5 presents empirical results and key findings. Section 6 discusses implications, conclusions, and recommendations. Finally, Section 7 highlights limitations and potential avenues for future research.

2 LITERATURE REVIEW

Student loans have been a growing concern over the past three decades, with a 7.2% or higher annual growth rate in related academic publications since 1990 (Bhandary and Ghosh, 2025), with an accelerated surge after 2020, likely driven by COVID-19-related forbearance programs. Comparing 2024 data with pandemic-era patterns shows that about 40% of borrowers made payments, and 30% missed them—a pattern similar to the 27% delinquency rate in February 2020 (Conkling and Gibbs, 2020). Recent studies have expanded beyond loan structures to encompass the emotional and psychological burdens of student debt, as noted by Bhandary et al. (2024), which highlights the mental health impacts associated with student borrowing. Sinha et al. (2024) discuss how student loans can delay life milestones such as marriage or homeownership, increasing anxiety, depression, and other disorders. Dwyer et al. (2012) report that borrowers with debt of less than \$10,000 were more likely to complete their degrees, whereas higher balances were associated with lower graduation rates. With the average borrower owing \$28,950, (Hahn, 2024) many students face debt levels that can significantly influence both academic outcomes and post-graduation life decisions.

Given these challenges, questions arise about how borrowers are expected to repay these increasing debts. Although several Income-Driven Repayment (IDR) plans exist, including SAVE, PAYE, IBR, and ICR (Willeford, 2025), Goldstein et al. (2023) argue that these plans often disadvantage low-income borrowers due to high administrative burdens, limited accessibility, and inconsistent implementation. They highlight the need for research that examines how policy design and administrative failure affect borrower outcomes, an area where our study contributes by applying statistical and ML-based insights into repayment behavior. Several studies have investigated the factors that influence repayment rates: Luo et al. (2018) identified student demographics (“Student”), institutional completion rates (“Completion”), and financial aid received (“Aid”) as key drivers of repayment success. Their study employed Principal Component Analysis (PCA) with a linear regression model, achieving an R^2 of 0.842 and an RMSE of 0.0197. Their RF Regression further reduced RMSE to 0.0153, identifying family income (FAMINC) as the most influential predictor of loan repayment, and the Pell Grant aid was also highly significant, indicating that it improves repayment outcomes. Similarly, Brown et al. (2019) analyzed student-loan-level data and emphasized the influence of institution type and degree category on repayment behavior. Notably, Luo et al. (2018) found no significant year-to-year variation in repayment rates between 2007 and 2013—an observation made using boxplots, paralleling the temporal trend analysis in our own study. Collectively, their findings validate the importance of institutional, demographic, and financial factors in repayment prediction, aligning closely with our ML approaches.

Beyond statistical modeling, some researchers focus on structural solutions to address repayment challenges. Boutros et al. (2024) propose two alternatives: Principal Payment Deferral (PPD), which allows deferral of principal payments while requiring continued interest payments, and

Full Payment Deferral (FPD), which permits temporary suspension of both principal and interest payments—an option particularly useful for recent graduates who are unemployed or underemployed. Financial literacy also plays a vital role in repayment success. Brown et al. (2019) found that students with stronger backgrounds in math and financial education were more likely to manage their loans effectively. Additionally, generational wealth and need-based support, such as Pell Grants, further reduce the risk of default. While not all studies include this factor explicitly, emerging data suggest that economic background is one of the most powerful predictors of long-term repayment success (Dearden 2019).

Policy changes and administrative decisions also shape repayment behavior. During the COVID-19 pandemic, the government paused federal loan repayments, a policy that ended in October 2023 (Bhandary and Ghosh, 2025). The Biden presidential administration forgave \$441 billion in federal student loans, but this initiative was blocked by the Supreme Court (Hahn, 2024). The Trump administration (as of 2025) has introduced new repayment plans effective July 1, 2026, structured similarly to mortgage repayments, which would be non-dischargeable (Turner, 2025). These changes will not affect current borrowers, who can continue to enroll in income-driven repayment plans. However, executive actions—including attempts to eliminate the Department of Education—signal further uncertainty in the student loan landscape (Dunbar, 2025).

While prior studies have identified key predictors of repayment rates and explored structural policy solutions, fewer have used machine learning to examine repayment across distinct borrower subgroups. Algorithmic models, such as Random Forest (RF), eXtreme Gradient Booster (XGBoost), and Nu Support Vector Regressor (NuSVR), can reveal nonlinear relationships and hidden patterns that traditional statistical methods may miss. This study aims to fill that gap by applying one-way Analysis of Variance (ANOVA) to assess significant yearly differences in repayment rates and ensemble ML models to evaluate repayment prediction performance across institutional and demographic subgroups, offering a more robust empirical foundation for more effective policymaking and interventions for student borrowers.

2 DATASET DESCRIPTION

The primary dataset used in this study is sourced from the U.S. Department of Education College Scoreboard (2025). This dataset provides comprehensive information at both the institution level and by field of study, covering data since 1996. It includes detailed statistics on student completion, debt and repayment, earnings, and other relevant information. The institution-level data aggregates information for each educational institution, including variables such as enrollment, student aid, costs, and student outcomes. The field of study includes information at the credential level, using 4-digit Classification of Instructional Programs (CIP) codes, and contains variables like cumulative debt at graduation and earnings one year after graduation.

For our analysis, we utilized the institution-level datasets from 2013 to 2017. This subset offers statistics and descriptions for thousands of U.S. educational institutions and is among the most recent datasets with minimal missing data. The original dataset comprises detailed data on 7,238 public universities, covering variables such as Unit ID for the institution, average Stafford and Grad PLUS loan debt disbursed at this institution, federal student loan borrower-based 2-year borrower count of completers, average cost of attendance (academic year institutions), percent of male/female students who completed within 6 years at the original institution, percent completed

within 6 years at the original institution, etc. Given the dataset’s large size and high dimensionality, we conducted thorough exploration and preprocessing. Records with missing values were excluded, resulting in a final subset of 1,062 public universities. From these, we selected 18 potentially important features for our study. Out of those, 15 were numeric variables, and 3 were categorical variables. The variables that are taken under study are: year, repayment rate, percentage of undergraduate students who completed within 6 years at the original institution, percentage of low-income, middle-income, and high-income students, female, male, students who received Pell grant, who never received a Pell grant, who received a federal loan, first generation students, non-first generation students who completed within 6 years at the original institution. Other variables being studied are the average cost of attending academic year institutions and the average net price for Title IV public institutions. The categorical variables under study are the institution’s name, unit ID, and control status.

Table 1 lists the main variables considered in this study. Additionally, Figure 1 illustrates the temporal trends of key student and institutional variables from the academic years 2013–2014 to 2016–2017. The variables include the 7-year repayment rate (RPY_7YR_RT) and 6-year completion rates by income level, gender, and grant/loan status. Each line represents a distinct academic year, enabling comparative analysis over time. Figure 2, the boxplot titled "Repayment Rate by Year" illustrates the distribution of student loan repayment rates across four academic years: 2013–2014, 2014–2015, 2015–2016, and 2016–2017. Similarly, Figure 3 is a correlation heatmap that provides an overview of the linear relationship between each pair of variables in the study, with a focus on the student loan repayment rate (RPY_7YR_RT).

TABLE 1. Variables and Features Used in Analysis

Variables	Features
Repayment Rate (RPY_7YR_RT)	Fraction of repayment cohort who are not in default and with loan balances reduced for seven years since entering repayment
Completion rate (COMP_ORIG_YR6_RT)	Percent of students who completed within 6 years at the original institution
Low Income (LO_INC_COMP_ORIG_YR6_RT)	Percent of low-income students (family income < \$30k)
Middle Income (MD_INC_COMP_ORIG_YR6_RT)	Percent of middle-income students (family income \$30k–\$75k)
High Income (HI_INC_COMP_ORIG_YR6_RT)	Percent of high-income students (family income > \$75k)
Female (FEMALE_COMP_ORIG_YR6_RT)	Percent of female students
Male (MALE_COMP_ORIG_YR6_RT)	Percent of male students
Pell Grant University (PELL_COMP_ORIG_YR6_RT)	Percent of students who received a Pell Grant at the institution
No Pell Grant (NOPELL_COMP_ORIG_YR6_RT)	Percent of students who never received a Pell Grant at the institution
Federal Loan (LOAN_COMP_ORIG_YR6_RT)	Percent of students who received a federal loan at the institution

First Generation (FIRSTGEN_COMP_ORIG_YR6_RT)	Percent of first-generation students
Not First Generation (NOT1STGEN_COMP_ORIG_YR6_RT)	Percent of students who are not first-generation
Price (NPT4_PUB)	Average net price for Title IV institutions
Cost (COSTT4_A)	Average cost of attendance
Pell Grant Undergraduate (PCTPELL)	Percentage of undergraduates who receive a Pell Grant

Note: All variables used in this study pertain to students who completed their studies within six years at their original institution.

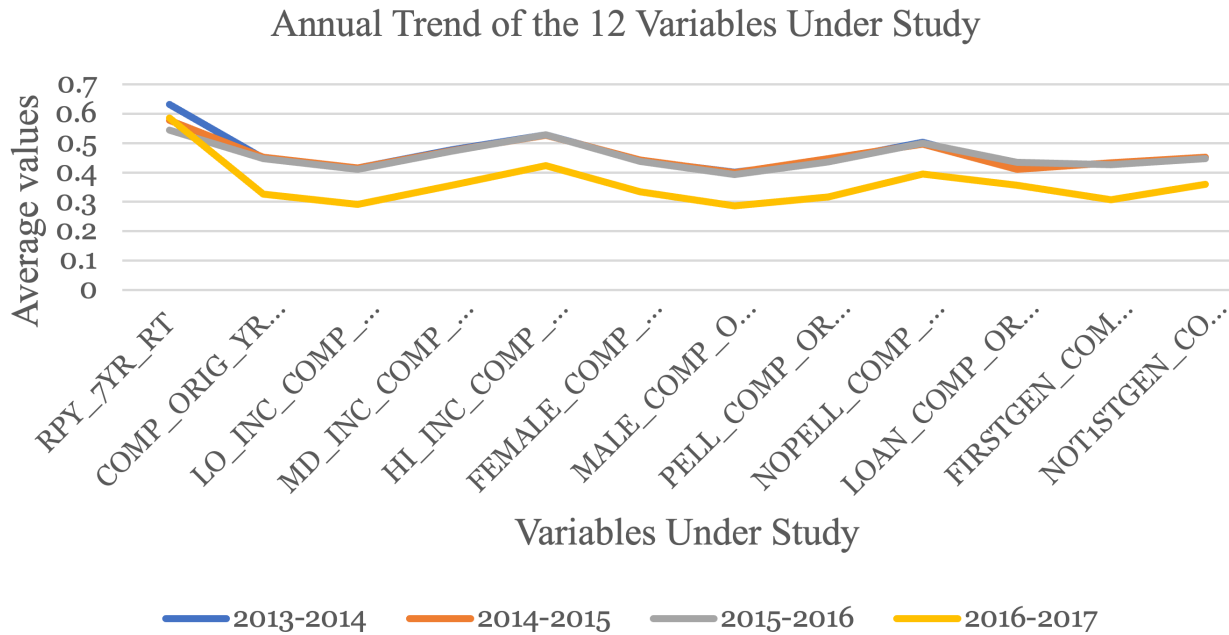


Figure 1. Average annual trends for the 12 variables included in this study, shown for the years 2013–2017.

4 METHODOLOGIES

4.1 Data preprocessing

The original dataset spanning the years 2013 to 2017 was sourced from the U.S. Department of Education College Scorecard (2025). Upon initial inspection, several inconsistencies were detected within the variables across the years. To ensure data quality, we performed thorough cleaning by removing missing values, null entries, and cells marked as “Privacy Suppressed.” After cleaning each year’s dataset individually, we merged all yearly datasets into a single unified dataset, ordered chronologically starting from the most recent year. The consolidated dataset was then split into training (80%) and testing (20%) subsets, using a random state of 100 and shuffle=True, to facilitate the development and evaluation of the machine learning model.

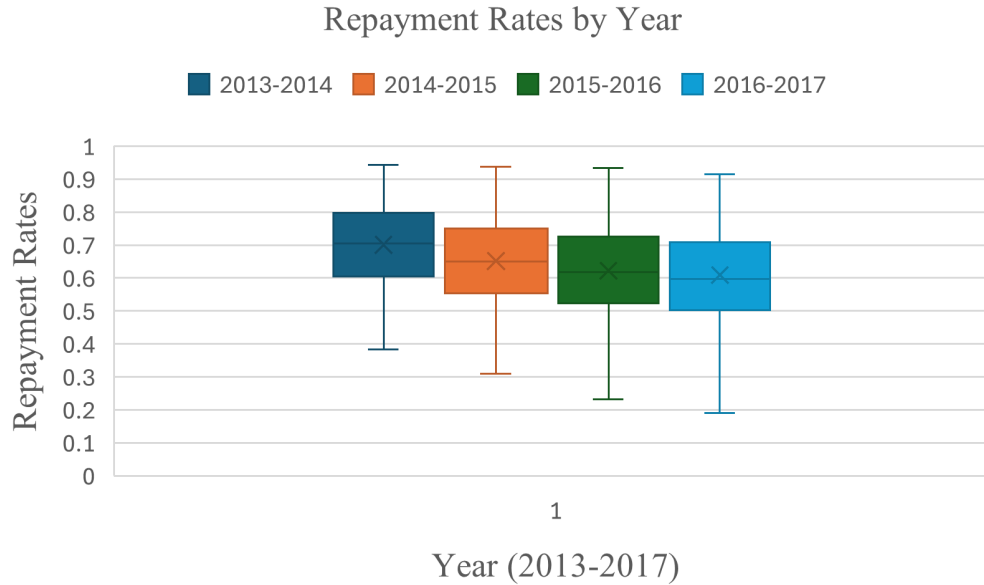


Figure 2. Yearly fraction of repayment cohorts who are not in default and whose loan balances have declined seven years since entering repayment, spanning the years 2013 to 2017.

4.2 Machine learning models

We used various ML models in our work to evaluate predictive performance for loan repayment rates. We used seven widely used ML regression methods and one ensemble method.

4.2.1 Model 1: Random Forest (RF): A random forest is a commonly used ML algorithm for classification and regression tasks. This method uses a variety of decision trees to make predictions. Each tree looks at different parts of the data randomly. For classification tasks, the result of the random forest is the class selected by the most trees. Whereas, for the regression tasks, the output is the aggregate of the predictions of the trees (Ho, 1995). It is a well-rounded method because it averages out individual tree errors, making the final prediction more accurate (Shalev-Shwartz and Ben-David, 2014). The hyperparameter configuration for this experiment was ($n_estimators=100$).

4.2.2 Model 2: Gradient Boosting (GB): Boehmke and Greenwell (2019) describe Gradient Boosting as an ensemble method of shallow trees in sequence; every tree after one is constructed is built to correct the error of the prior ones. The first tree built is called the “weak model”; the next tree is then constructed on the residual error, effectively and gradually refining the accuracy.

4.2.3 Model 3: Decision Tree (DT): A Decision Tree is a predictor tree where a step-by-step decision process occurs (Shalev-Shwartz and Ben-David, 2014). They explain that the model classifies an input by traversing from a root to a leaf via successive internal node tests based on feature values, ultimately returning the label stored in the leaf that is reached. The specific path taken depends on the value of the feature in the input.

4.2.4 Model 4: Bagging Regressor (BR): Boehmke and Greenwell (2019) describe Bagging Regressor as a method that uses an ensemble of fitting multiple models to different bootstrapped samples of the data and then averages the predictions for classification. The process helps reduce the variance and is especially helpful when using high-variance models. The hyperparameter configuration for this experiment was ($n_estimators=70$).

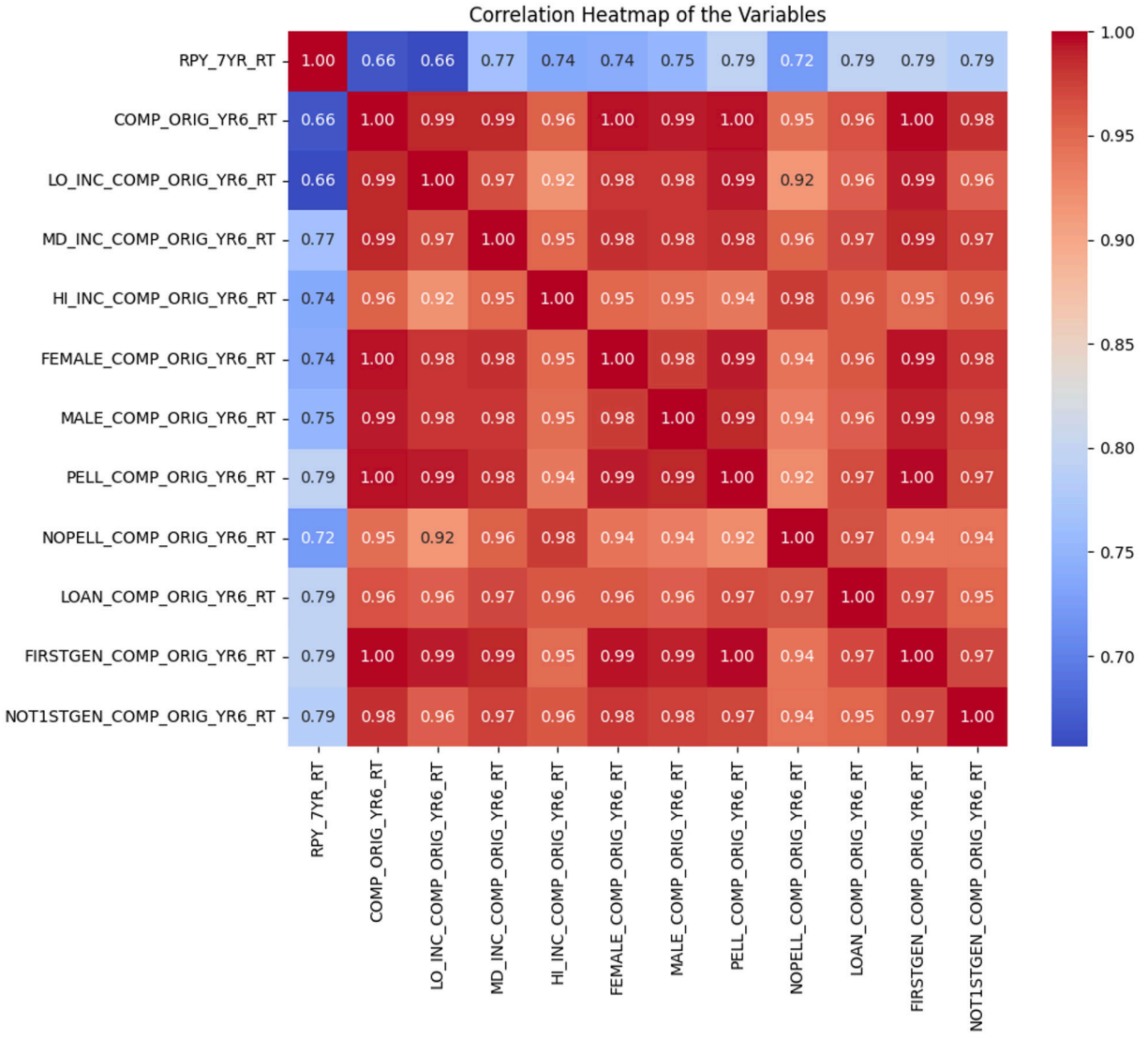


Figure 3. The heatmap matrix illustrates the correlation coefficients among the key features used in this study. This visualization illustrates the strength and direction of linear relationships between variables that are relevant to predicting student loan repayment.

4.2.5 Model 5: K-Nearest Neighbors (KNN): KNN is an algorithm that uses the closest training examples and predicts based on their labels, either by averaging for regression or by the majority vote for classification (Shalev-Shwartz and Ben-David, 2014). The labeled distance is measured using a distance function, usually referred to as Euclidean distance. In our experiment, the parameters we used were 3 neighbors ($n_neighbors = 3$) and a leaf size of 30 ($leaf_size = 30$). The k-dimensional tree (KDTree) algorithm was used to compute nearest neighbors ($algorithm = kd_tree$) with a distance-weighted function ($weights = distance$).

4.2.6 Model 6: eXtreme Gradient Booster Regressor (XGB): Boehmke and Greenwell (2019) highlight XGB as a model that serves as an improved gradient boosting framework built for performance, adaptability, and portability across various programming languages. The authors build on this definition by highlighting other advantages that this model has over other traditional boosting methods, including early stopping, saving results with the existing model, parallel processing,

regularization, and much more. XGB builds on gradient boosting, but it has certain enhancements that make it more effective and functional. In our experiment, we used 1000 boosting rounds ($n_estimators = 1000$), with a 0.1 step size shrinkage ($\eta = 0.1$), and a maximum depth of 5 for each tree ($max_depth = 5$).

4.2.7 Model 7: Nu Support Vector Regressor (NuSVR): Langhammer and Česák (2016) explain that the “Nu” in NuSVR is a parameter that controls the fraction of training errors and the number of support vectors used. Montesinos-Lopez et al. (2022) note that the “SVR” stands for Support Vector Regression, a method that utilizes residuals larger than a certain amount, known as epsilon. Like Support Vector Machines (SVMs) for classification, SVR penalizes points outside a “tube” around the prediction line. Montesinos-Lopez et al. (2022) explain later that this approach finds the best-fitting line in a transformed feature space so it can generalize well to new data.

4.2.8 Model 8: Stacking Ensemble methods: Dietterich (2000) explains that in machine learning, ensemble methods consist of techniques that build a set of classifiers and then combine their outputs—typically via weighted or unweighted voting to make certain predictions. He notes that these two methods outperform any individual classifier because they possess two key properties: accuracy and diversity. There are several ensemble methods in existence, with the original one being Bayesian averaging, and more recent methods including Bagging and Boosting.

TABLE 2. Summaries of Calculated RMSE and R² Values for Evaluated Methods

Model	RMSE	R ²
RF	0.067	0.7607
DT	0.093	0.5393
GB	0.071	0.7273
BR	0.067	0.7600
KNN	0.070	0.7415
XGB	0.066	0.7667
NuSVR	0.063	0.7900
Ensemble model	0.0038	0.7962

Performance analysis of the evaluated machine learning models showing the RMSE and R² scores of 4-year data on the test dataset. The stacking ensemble model demonstrates the best overall predictive accuracy among all methods.

For our experiments, we employed stacking ensemble methods, utilizing three models (RF, XGB, and NuSVR) with high performance as the first-level base models. The base model is fitted using 5-fold cross-validation ($random\ state=0$, $n_folds=5$, $shuffle=True$), and after fitting the entire dataset, predictions are made on the testing data. We have utilized the NuSVR as a second-level model (meta-model) for our experiments as from our preliminary experiments, NuSVR shows the highest R² value as shown in Table 2. The train stacking features from the first-level models are used to train the model, which is used to predict the final output with test stacking features. Figure 4 illustrates the block diagram of stacking ensemble methods, utilizing three models as base models.

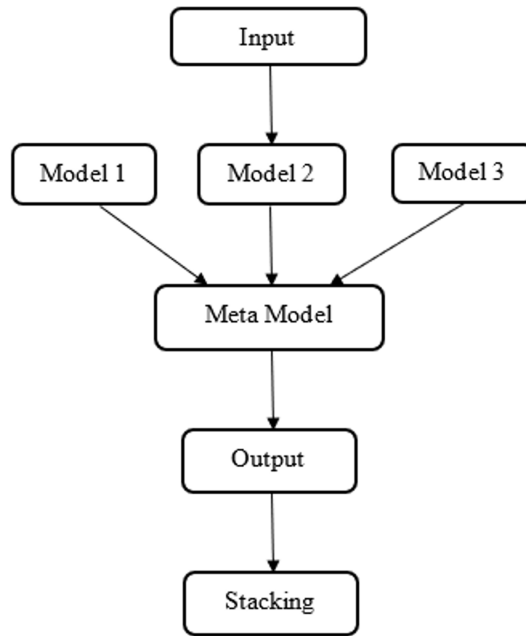


Figure 4. Block diagram of the stacking ensemble model. The first level consists of three base models—RF, XGB, and NuSVR—each trained using 5-fold cross-validation. Their predictions are combined with the creation of stacking features, which are used to train the second-level meta model. The meta model produces the final prediction.

4.3 Performance evaluation metrics

For our evaluation of the models, various performance metrics were used to assess the effectiveness of the models in predicting student loan repayment. We used two commonly used evaluation metrics for predicting student loan repayment: the Root Mean Squared Error (RMSE) score and the R-squared (R^2) score to measure performance on the testing data. Both analyze the difference between the target value and the actual value.

RMSE metrics measure the average magnitude of prediction errors, quantifying the typical difference between the predicted and actual values in the regression model. It is calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (1)$$

where y_i is the targeted value, \hat{y}_i is the actual value for the i^{th} observation, and n is the sample size.

R^2 indicates the proportion of the variance in the dependent variable that is explained by the independent variables in the regression model. It is calculated as:

$$R^2 = 1 - \frac{SSR}{SST} \quad (2)$$

$$= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

where SSR is the Sum of Squares of the residuals, SST is the Sum of Squares of the Total, \bar{y} is the mean of the actual y values.

5 EXPERIMENTAL RESULTS

One of our primary goals was to study the trend of the variable Repayment rate over four years and to predict the future student loan repayment rates. We began by analyzing the repayment rates from these four years using a one-way ANOVA performed in R statistical software (version R-4.5.1). The objective was to determine whether there were statistically significant differences in repayment rates across the years, with significance set at $p \leq 0.05$.

Further pairwise comparisons using the Tukey-Kramer method revealed that the most notable differences occurred between the years 2015–2016 and 2016–2017, suggesting a potential shift in repayment behavior during this period.

Next, we evaluated the predictive performance of seven different ML models. Among these, NuSVR demonstrated the highest performance with an R^2 score of 0.79 and an RMSE of 0.063, outperforming the other models. The RF and XGB models yielded similar strong results, with R^2 scores of 0.7607 and 0.7667, respectively, and RMSE scores of 0.067 and 0.066. The BR model also performed comparably, with an R^2 of 0.76 and an RMSE of 0.067. In contrast, the DT model showed the weakest performance, with an R^2 value of 0.5393, indicating a weaker correlation between the predicted and actual repayment rates.

Based on these evaluations, the three best-performing models (NuSVR, RF, and XGB) were selected for further analysis using stacking ensemble methods. The stacking ensemble showed a slight improvement, achieving an R^2 score of 0.7962 and an RMSE of 0.0038. Figure 5 represents scatter plots comparing the actual repayment rates against predictions from the three individual models and the ensemble method. The plots show data points concentrated closely along the regression line with only a few outliers, indicating a strong correlation between actual and predicted values and suggesting that these models provide reliable predictions. Table 2 summarizes the calculated RMSE and R^2 values for each method evaluated.

6 DISCUSSION AND CONCLUSION

This study aimed to investigate the key predictors of federal student loan repayment and assess whether ML models can improve prediction accuracy. Through statistical and machine learning analyses, we found that specific demographic, financial, and institutional factors have a significant influence on student loan repayment behaviors. Our findings provide several important insights for policymakers and administrators seeking to design evidence-based and equitable loan support strategies, as well as for predictive modeling purposes.

The performance of various machine learning models in predicting the 7-year repayment rate was evaluated, with NuSVR emerging as the top performer among the individual models. It achieved R^2 of 0.79 and RMSE of 0.063, outperforming both RF and XGB, which also demonstrated strong results. The ability of NuSVR to explain a greater variance in the repayment rate suggests that nonlinear modeling approaches are better suited to capture the complex financial and institutional factors that affect borrowers' outcomes. From a policy standpoint, these methods and levels of precision enable policymakers to accurately identify students at risk, support earlier interventions, more effectively target financial support resources, and provide more personalized repayment guidance.

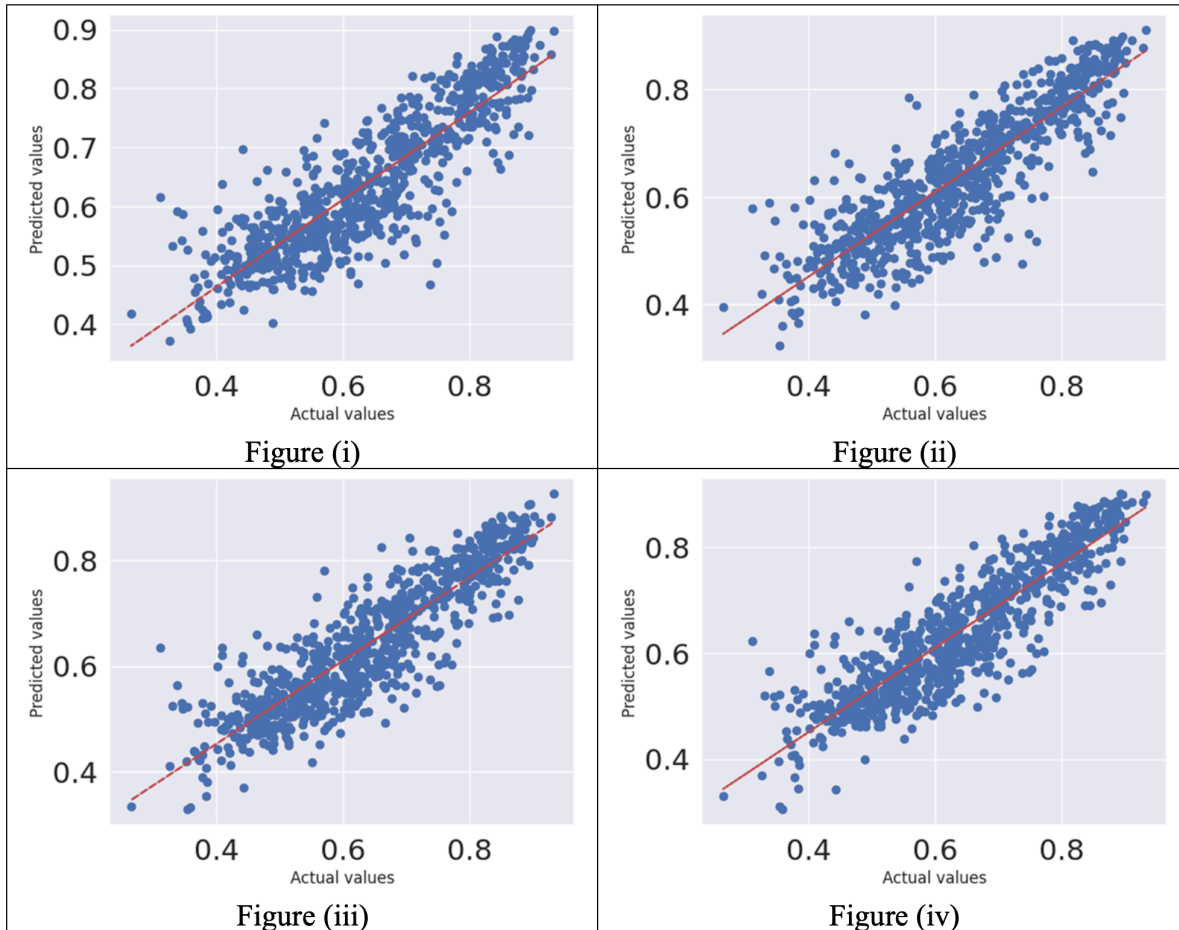


Figure 5. Scatter plots comparing actual versus predicted repayment rates for (i) RF, (ii) XGB, (iii) NuSVR, and (iv) the stacking ensemble model. Points are clustered closely around the regression line, indicating strong predictive performance and reliable model fits. Outliers are minimal across all models.

Our stacking ensemble model, which combined the predictions of NuSVR, RF, and XGB with NuSVR as the meta-learner, further improved performance, achieving an R^2 of 0.7962 and an RMSE of 0.0038. This improvement highlights the importance of combining multiple modeling approaches to produce more reliable repayment predictions. For policymakers and higher education leaders, it helps in accurate forecasting with the insights needed for evidence-based decision-making. These evidence-based decisions could enhance financial loan counseling programs and implement institutional-level accountability measures. Overall, these results demonstrate the potential of how advanced predictive analytics can guide the development of more precise, proactive, and equity-oriented student loan policies.

Statistical analysis using ANOVA and post hoc Tukey-Kramer comparisons revealed significant yearly differences in repayment rates, particularly between the 2015–2016 and 2016–2017 periods. These differences coincide with significant policy and economic changes during this period, including alterations in federal loan programs and shifts in institutional aid, suggesting that such external factors may influence borrower behavior. Additionally, feature importance analysis and correlation matrices revealed that completion rate, family income levels, gender, and Pell Grant status were key factors influencing repayment outcomes. These findings are consistent with the

prior research [e.g., (Luo et al., 2018); (Brown et al., 2019)] and reinforce the critical role of student background and financial support in shaping repayment behavior. These results, which link repayment trends to specific policy and economic contexts, provide practical guidance for policymakers to address factors affecting loan repayment.

Overall, our results suggest that machine learning models, especially ensemble approaches, are powerful tools for predicting student loan repayment outcomes. The integration of institutional, demographic, and financial features provides a comprehensive framework for understanding borrower behavior. By identifying key predictors and accurately forecasting repayment success, this research can help policymakers, educational institutions, and financial aid offices design more effective support systems and repayment plans. Given the increasing student loan burden in the United States, now exceeding \$1.7 trillion, such insights are critical. While repayment plans and forgiveness programs have aimed to reduce financial pressure, prediction models like those presented in this study can further help target interventions, improve loan counseling, and allocate resources more effectively.

7 LIMITATIONS AND FUTURE WORK

While the results are promising, this study has several limitations. The dataset only includes public institutions and students who completed their degrees in six years, which limits the generalizability of the findings. The analysis was conducted on institution-level data, which may have masked significant variability in student experiences and outcomes. Additionally, the study examined only a limited range of features and machine learning models, which may have constrained the robustness of the findings.

Future research directions include expanding the dataset to the most recent years, incorporating private institutions and students with varied completion timelines, and including a broader and more granular set of socioeconomic and academic variables. Beyond research needs, these findings point towards the actionable steps for policymakers and higher education leaders. From a research perspective, investments in comprehensive federal data systems that link institutional characteristics, financial aid records, post-graduation outcomes, and targeted loan repayment interventions would enable more precise risk modeling and causal analysis. Such integrated data would allow researchers to better identify patterns of loan distress and evaluate the effectiveness of repayment policies. In parallel, institutions could use research-informed predictive models to identify at-risk students earlier and test evidence-based interventions, including enhanced academic advising, financial counseling, and repayment-planning supports, to mitigate the structural and behavioral factors associated with student loan distress. Future studies may employ advanced modeling techniques, such as deep learning and time-series analysis, to more accurately capture long-term repayment dynamics.

Despite these limitations, this study demonstrates the potential of data science and machine learning to make meaningful contributions to understanding and addressing one of the most pressing financial challenges facing American students today: student loan repayment.

AUTHOR INFORMATION

Corresponding Author*

Prabha Shrestha
Statistics and Mathematical Sciences
Eastern New Mexico University
Portales, New Mexico 88130
prabha.shrestha@enmu.edu

REFERENCES

- 10 Key Facts about Student Debt in the United States*. Peter G. Peterson Foundation, 2024. <https://www.pgpf.org/article/10-key-facts-about-student-debt-in-the-united-states/>
- Bhandary, R.; Ghosh, B. K. Credit card default prediction: An empirical analysis on predictive performance using statistical and machine learning methods. *Journal of Risk and Financial Management* **2025**, *18*(1), 23. DOI: 10.3390/jrfm18010023
- Bhandary, R.; Shenoy, S. S.; Shetty, A.; Shetty, A. D. Education loan repayment: A systematic literature review. *Journal of Financial Services Marketing* **2024**, *29*(4), 1365–1376.
- Boehmke, B.; Greenwell, B. M. *Hands-on Machine Learning with R*; CRC Press, 2019.
- Boutros, M.; Clara, N.; Gomes, F. Borrow now, pay even later: A quantitative analysis of student debt payment plans. *Journal of Financial Economics* **2024**, *159*, DOI: 10.1016/j.feineco.2024.103898.
- Brown, M.; Chakrabarti, R.; der Klaauw, W.; Zafar, B. Understanding the evolution of student loan balances and repayment behavior: Do institution type and degree matter? *Economic Policy Review* **2019**, *25*(1), 1–23.
- College Scoreboard, U.S. Department of Education. <https://collegescorecard.ed.gov/data>.
- Conkling, T. S.; Gibbs, C. *An Analysis of the First Seven Months of the Federal Student Loan Return to Repayment*; Consumer Financial Protection Bureau Office of Research Reports Series, 2024, No. 24-6.
- Dearden, L. Evaluating and designing student loan systems: An overview of empirical approaches. *Economics of Education Review* **2019**, *71*, 49–64.
- Diaz, M. Mental Health Impact of Student Loan Debt on College Students. M.S. Thesis, California State University, San Bernadino, 2025.
- Dietterich, T. G. Ensemble methods in machine learning. *Multiple Classifier Systems*; 2000; pp 1–15.
- Dunbar, M. US student loan collections resume: Here’s what you need to know. *The Guardian*, May 6, 2025. <https://www.theguardian.com/us-news/2025/may/06/us-student-loan-collections-explainer>.

- Dwyer, R. E.; McCloud, L.; Hodson, R. Debt and graduation from American universities. *Social Forces* **2012**, *90*(4), 1133–1155.
- Edwards, D. How we can solve the student loan debt crisis. *The Journal of the James Madison Institute*. **2016**, *Winter*, 79-90.
- Goldstein, A.; Eaton, C.; Villalobos, A.; Chakrabarti, P.; Cohen, J.; Donnelly, K. Administrative burden in federal student loan repayment, and socially stratified access to income-driven repayment plans. *The Russell Sage Foundation Journal of the Social Sciences* **2023**, *9*(4), 86–111.
- Hahn, A. Student loan debt statistics: Average student loan debt. *Forbes*, April 18, 2024. <https://www.forbes.com/advisor/student-loans/average-student-loan-debt-statistics/>
- Hegji, A.; Stiff, S. M. *The Biden Administration's Student Loan Debt Relief Rulemaking*; Congressional Research Service, 2024. <https://www.congress.gov/crs-product/IN12350>
- Heller, D. E. The impact of student loans on college access. *The Effectiveness of Student Aid Policies: What the Research Tells Us*; Baum, S.; McPherson, M.; Steele, P., Eds.; College Board, **2008**; pp 39–68.
- Ho, T. K. Random Decision Forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, Canada, August 14-16, 1995; Vol. 1, pp 278–282. DOI:10.1109/ICDAR.1995.598994.
- Jabbari, J.; Roll, S.; Despard, M.; Hamilton, L. Student debt forgiveness and economic stability, social mobility, and quality-of-life decisions: Results from a survey experiment. *Socius* **2023**, *9*, DOI: 10.1177/23780231231196778.
- Kuan, R.; Blagg, K.; Castleman, B. L.; Darolia, R.; Matsudaira, J. D.; Milkman, K. L.; et al. Behavioral nudges prevent loan delinquencies at scale: A 13-million-person field experiment. *Proceedings of the National Academy of Sciences* **2025**, *122*(4), DOI: 10.1073/pnas.2416708122.
- Langhammer, J.; Česák, J. Applicability of a nu-support vector regression model for the completion of missing data in hydrological time series. *Water* **2016**, *8*(12), 560.
- Luo, B.; Zhang, Q.; Mohanty, S. D. Data-Driven Exploration of Factors Affecting Federal Student Loan Repayment. *arXiv preprint*, May 3, 2018. DOI:10.48550/arXiv.1805.01586.
- Melton, M. Financial Literacy in Higher Education: A Study on College Graduates Between 2012–2022 and Their Financial Thought Process on Selecting Their College, Payment Plan, Major, and Financial Repercussions. D.Ed. Dissertation, Union University, 2023.
- Montesinos López, O. A.; Montesinos López, A.; Crossa, J. Support vector machines and support vector regression. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*; Springer, 2022; pp 337–378.
- Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press, 2014.

Sinha, G. R.; Viswanathan, M.; Larrison, C. R. Student loan debt and mental health: a comprehensive review of scholarly literature from 1900 to 2019. *Journal of Evidence-Based Social Work* **2024**, *21*(3), 363–393.

Turner, C. The Future of Student Loan Repayment, Explained. *National Public Radio*, May 12, 2025. <https://www.npr.org/2025/05/12/nx-s1-5389644/trump-student-loan-program-forgiveness-overhaul>

Willeford, J. SCHOOL'S OUT Students to see simplified loan repayment plan in new July 1 proposal—two options 'address root cause' of high prices. *The U.S. Sun*, April 29, 2025. <https://www.the-sun.com/money/14130364/students-simplified-loan-repayment-gop-proposal>